ChinaXiv合作期刊

Springer Nature SciGraph 关联开放数据分析*

白林林[1,2] 祝忠明[1]
[1] 中国科学院兰州文献情报中心 兰州 730000
[2] 中国科学院大学 北京 100049

摘要： ［目的/意义］Springer Nature SciGraph 平台提供的关联开放数据的分析，为国内出版商在学术交流和语义出版中使用关联数据促进科研的关联开放实践提供参考，可推动我国开放科研运动的进一步向前发展。[方法/过程] 对Springer Nature SciGraph 平台发布的实体对象、采用的词表、数据模型进行详细分析。[结果/结论] Springer Nature SciGraph 通过构建自己的本体，采用一种用于 RDF 的更简单的序列化 N-Triples 格式的三元组对数据进行表示，作为世界上最大的出版商之一，Springer Nature 关联数据在今后必将为其他出版商要实现科研关联开放提供一定的借鉴。

关键词： Springer Nature SciGraph 关联开放数据开放科研
分类号： G254
基金项目： 本文系中国科学院文献情报能力建设专项"中国科学院知识资产存缴管理中心建设"项目（项目编号：Y6ZG421001）研究成果之一。

作者简介： 白林林(ORCID:0000-0003-2265-7399)，博士研究生，E-mail：bailinlin@mail.las.ac.cn；
祝忠明（ORCID：0000-0002-2365-3050），信息系统部主任，研究员。

1 引言

关联数据作为一种最佳的语义网实践，从 2006 年提出至今十多年的时间，经历了从最简单的知识库与词汇表到不用领域的应用。L. Jens、A. Sören、C. Sarven 等在 2017 年第 10 届 LDOW（Linked data on the Web）会议上对关联数据在过去十年的的发展进行了总结，并指出在未来的十年，关联数据在学术交流中将发挥较大的作用[1]，且此次会议提出“开创关联开放科研知识云（pioneering the linked open research cloud）”[2]的新型倡议，鼓励学者将关联数据技术与最佳实践应用于学术交流中。C. Sarven 等提出关联开放原则（linked research priciple）以鼓励关联科研知识的开放[3]。ScholarlyData.org[4]利用会议本体数据模型[5]对语义网会议相关的论文、人员、组织和事件进行了组织。欧盟、美国政府和澳大利亚政府等在 2013 年成立国际研究数据联盟（Research Data Alliance）[6]，利用数据标准和实践促进科研数据的共享与交换，通过科研数据转换平台（research data switchboard）与科研图谱（research graph）项目实现了连接科研人员、出版物、研究资助和研究数据集（研究数据）的功能。Springer Nature 为帮助科研共同体充分利用开放科研所带来的便利，新推出 SciGraph 关联开放数据平台[7]，集成了 Springer Nature 及其合作伙伴的数据资源，如有关科研资助机构、科研项目及拨款、会议、科研单位和出版物的信息，让分析 Springer Nature 出版物的相关信息变得更加容易，目前，该平台包含了 1.55 亿条学术界关注对象的信息（三元组）。更多的数据，如引用、专利、临床试验和使用数量等，也将分阶段推出，这样到 2017 年底，Springer Nature SciGraph 的三元组数量将增至 10 亿条以上[8-9]。

本文以 Springer Nature SciGraph 平台提供的关联开放数据为研究对象，对其发布的实体对象、采用的本体和词表、数据模型进行分析，旨在为国内在学术交流和开放科研中使用关联数据提供参考，促进国内科研数据的语义化、国际化共享和开放科研运动的进一步

chinaXiv:201801.00105v1

向前发展。

2 Springer Nature SciGraph 的实体对象分析

Springer Nature 的关联开放数据中分了主体（agent）、资产（asset）、概念（concept）、事件（event）四大类，其中概念和事件类中包括的实体是 Springer Nature 发布的重点。概念的下一层分类包括：注释、合同、出版物、类型，注释的子类主要是引文计量类，合同的子类是资助类，出版物的子类包括产品（产品指 Springer Nature 对外提供的文章、图书、书的章节、期刊）和作品（专著、连续出版物），类型包括主题、获取类型、文章类型、会议系列、产品市场编号、出版物状态子类。事件的下一层分类包括：隶属机构、聚合事件、注释事件、会议、贡献者、出版事件[8]。事物（thing）类是所有类的上位类。

在发布关联数据之前，需要明确待发布的数据中实体类型及实体间的关系。遵循关联数据发布原则第一条，即用 URI 作为任何资源的名称，确保资源的可获取性。Springer NatureSciGraph 的 URI 主要有两种模式：

http://www.springernature.com/scigraph/things/{datasets}/{scigraphId}

http://www.springernature.com/scigraph/things/{datasets}/{topic}

这两种模式是一样的，第二种模式主要是针对主题类实体对象的。首先以 http://www.springernature.com/scigraph/作为基地址，该地址作为 Springer NatureSciGraph 关联开放数据的发布平台。“thing”类是所有类的上位类。“datasets”为各类实体组成的数据集，其属性值可以有 articles、grants、journals、journalbrands、subjects、contributions、books 等，必须指定数据集才能访问其中的对象。对象集后必须有相应的对象，URI 中用“scigraphId”或“topic”表示，否则无法找到对象。

目前 Springer Nature 已发布的关联开放数据中实体类型有文章、期刊、主题、资助，各数据对应的 URI 如表 1 所示。

表 1 实体对象的 URI

	实体对象	URI
概念	article（文章）	http://www.springernature.com/scigraph/things/articles/{scigraphId}
	journal（期刊）	http://www.springernature.com/scigraph/things/journals/{scigraphId}
	subject（主题）	http://www.springernature.com/scigraph/things/subjects/{topic}
	grant（资助）	http://www.springernature.com/scigraph/things/grants/{scigraphId}

具体实例如：

主题 URI：http://www.springernature.com/scigraph/things/subjects/geology（在 Springer

chinaXiv:201801.00105v1

Nature 中的主题词 geology）

期刊 URI：

http://www.springernature.com/scigraph/things/journals/042783d5f9e6e3813522b5ebbe89f4ab

Springer Nature 不仅为数据建立了有效、唯一的 URI，还建立了一个 SciGraphcore ontology（前缀：sg）本体，这个本体由 45 个类和 206 个属性组成，拥有自己的命名空间（http://www.springernature.com/scigraph/ontologies/core/，前缀 sg :)。在概念上，这个本体是以前的 nature.com 核心本体[10]的延伸。构建这个本体的原因一是由于在其他本体或词汇表中找不到相应的词汇来描述某些数据或属性，二是符合本模型特色的类和属性，可以使 Springer Nature 在更好描述数据的同时被外界更好地引用。并与外部的全球研究标识符数据库（Global research identifier database, GRID）[11]、澳大利亚与新西兰标准研究分类法：研究领域（Australian and New Zealand standard research classification: fields of research，ANZSRC-FOR）[12]、DOI[13]建立了链接。

澳大利亚与新西兰标准研究分类法：研究领域（ANZSRC-FOR）是根据研发过程中所用的方法对研发活动进行分类，而不是根据研发单位或者研发目的进行分类。ANZSRC-FOR 分类法中的类别包括由企业、大学、高等学校、国立科研机构和其他组织研究探讨的主要研究领域及相关子领域和新型领域。

全球研究标识符数据库（Global research identifier database，GRID）不仅提供有关组织的 ID 和名称，而且提供了数据类型、等级结构、所处位置等元数据，与 GeoNames、WikiData、CrossRef、开放资助者注册表、国际标准名称标识符（international standard name identifier，ISNI）等实现链接，扩充了元数据。

3Springer Nature SciGraph 词表

关联数据的发布原则第三条是尽可能复用已有的、成熟的词表来描述资源，用以提高词汇表的互操作性，减少对本地词汇的管理。Springer Nature SciGraph 所用的词汇表主要分为通用词表和专用词表两类（如表 2 所示），其中通用词汇表主要用于描述实体的一般属性，如实体类型、实体类型之间的关系等；专用词表用以描述具体实体，并具有所描述实体的属性。可看出，Springer NatureSciGraph 用复用通用的词汇表来描述实体的类型、RDF 数据的元数据（VoID[14]）、注释所用的词汇等，使用的专用词汇表只有 SKOS 和自建的本体 SciGraph Core Ontology，并未复用其他词汇表对资源进行描述，但是 Springer Nature 提供了与 bibo[15]、crm（conceptual reference model）[16]、depedia[17]、depedia –owl、dc、dcterms、event、fabio（the FRBR-aligned bibliographic ontology）[18]、foaf、mesh（medical subject headings）[19]、obo（open biomedical ontologies）[20]、prism（publishing requirements for industry standard metadata）[21]、schema、skos、vcard[22]、vivo（integrated semantic framework）[23]、wd（wikidata）[24]

chinaXiv:201801.00105v1

chinaXiv:201801.00105v1

本体之间的映射，其中与 dbpedia、mesh、wd 是主题词之间的映射，其他是类和属性之间的映射。

可看出，SciGraph Core Ontology 所描述类和属性比较全面，虽然未复用已有成熟词表，但是使用根据自身需求建立的本体可以更准确地描述相关的类和属性。

表 2 词表及注释

词表		注释
通用词表	OWL（Web ontology language）	网络本体语言，用以在万维网中发布和共享本体
	RDF（resource description framework）	资源描述框架，一种通过“主-谓-宾”三元组形式描述 Web 资源的标记语言
	RDFS （ resource description framework schema）	是 RDF 词汇表的扩展词汇表，为 RDF 数据提供数据建模词汇表
	Dcterms（DCMI metadata terms）	都柏林词汇表的扩展词汇表
	DC（Dublin core）	都柏林词汇表
	Vann（a vocabulary for annotating vocabulary descriptions）	通过用例和使用说明注释词汇，描述进行注释的词汇表
	VoID（vocabulary of interlinked datasets）	用以对 RDF 数据的元数据进行描述
专用词表	SKOS （ simple knowledge organization System)	用以对受控词表中词汇进行描述的知识组织系统词汇表
	sg（SciGraph Core Ontology）	Springer Nature 自己构建的本体，用以描述 Springer Nature 网站提供的资源

4 Springer Nature SciGraph 数据模型分析

目前 Springer Nature 发布的关联开放数据中实体类型有文章、期刊、主题、资助者，这些数据之间的关系模型如图 1 所示[8]，文章通过 sg:hasJournal、sg:hasSubject 分别与期刊和主题进行链接，资助实体通过 sg:hasFundedPublication 与文章进行链接，如图 2 所示。本文将这个数据模型进行拆分并分别进行分析。

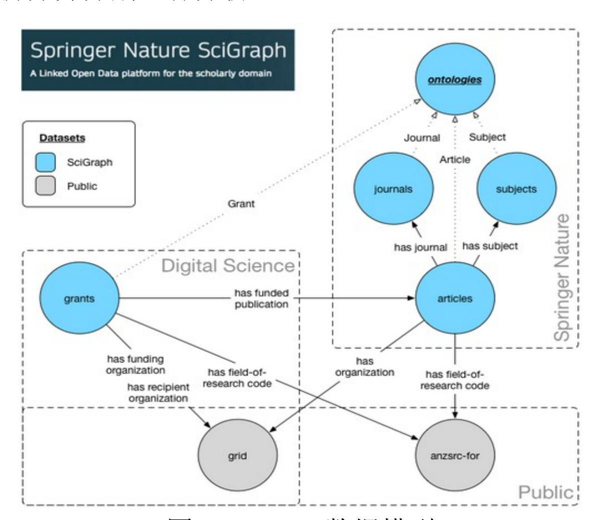


图 1SciGraph 数据模型

图 2 文章数据模型

4.1 文章数据模型

Springer Nature 为文章类数据模型提供了 20 个属性，将 20 个属性分为类型、标识符、标签、贡献者、出版商、主题信息、来源 6 类。使用 rdf:type 表示类型；标识符属性包括：scigraphId、数字对象标识符（digital object identifier，DOI）和数字对象标识符链接；标签属性包括：语种、题名、翻译题名、摘要、翻译摘要；出版商信息包括：出版年、出版年月、出版日期、Springer Nature 网页；主题信息包括澳大利亚与新西兰标准研究分类法：研究领域（ANZSRC-FOR）[11]和 Springer Nature 自己建立的主题词；来源信息包括期刊、期刊卷、期刊期号。

在各属性的取值中，与 Springer Nature 提供的贡献者、资助数据、期刊、主题词类对应的实例 URI 进行内部链接，选用数字对象标识符系统提供的 DOILINK 和 ANZSRC-FOR 提供的分类号作为外部链接，其他属性值取值为文本值或数值，见图 2。

Springer Nature 目前并没有贡献者（contribution）的数据模型。但是从其发布的 N-Triples[25]格式的三元组数据，可知贡献者属性有数据类型、scigraphId、对外公开的姓名、对外公开的姓、对外公开的名、排序（文章作者排序）、是否为通讯作者（属性值为布尔值 true 或 false）、角色（取值为“author”“editor”或“principal investigator”）、隶属机构（为 Springer Nature 提供的实体）[26]。

4.2 资助数据模型

Springer Nature 为资助类数据模型提供了 18 个属性，将 18 个属性分为类型、标识符、标签、资助额、资助时间、资助主体、其他 6 类。使用 rdf:type 表示类型；标识符属性为资助的 scigraphId；标签属性包括：语种、题名、翻译题名、摘要、翻译摘要；资助额信息包括：资助金额、融资货币；资助时间包括资助开始时间、资助结束时间；资助主体包括资助组织、被资助组织；其他属性包括与资助相关的贡献者、资助的研究领域分类号、资助的出版物、许可条款、所在网页。

在各属性的取值中，与 Springer Nature 提供的与资助相关的贡献者、资助的出版物文章对应的实例 URI 进行内部链接，选用全球研究标识符数据库（global research identifier database，GRID）[10]提供资助组织与被资助组织 URI 和 ANZSRC-FOR 提供的分类号作为外部链接，其他属性取值为文本值或数值，如图 3 所示：

chinaXiv:201801.00105v1

图 3 资助数据模型

4.3 期刊数据模型

Springer Nature 为期刊类数据模型提供了 8 个属性，将 8 个属性分为类型、标识符、所属期刊品牌、格式、是否为正在出版、是否为历史期刊 6 类。使用 rdf:type 表示类型；标识符属性包括为期刊的 scigraphId、ISSN 号、DOI；使用 sg:hasJournalBrand 表示所属期刊品牌；格式主要指期刊的媒介形式；是否为正在出版用 sg:isActivePublication 属性表示；是否有历史期刊用 sg:isHistoricalJourna 表示。

在各属性的取值中，与 Springer Nature 提供的与期刊所属的期刊品牌、出版在期刊上的文章对应的实例 URI 进行内部链接，其他属性取值为文本值或数值，其中期刊的媒介属性文本取值为“Electronic”和“Paper (journals, normal index)”，“是否为正在出版期刊”和“是否为历史期刊”属性取值均为为布尔值“true”或“false”，见图 4。

从 Springer Nature 提供的期刊数据模型的 N-Triples 格式的三元组数据可知，期刊品牌属性有数据类型、scigraphId、语种、标题、标题简称、副标题、版本说明、出版商、知识产权所有者、所在网页、添加 Springer Nature 数据库的日期、创建年代、结束年代、开始卷号、结束卷号、卷数、开放获取（取值为“Fully Open Access”或“Hybrid (Open Choice)”）。

图 4 期刊数据模型

4.4 主题数据模型

Springer Nature 的主题主要是自己网站建立的主题词，主要分为生物科学、地球与环境科学、生命科学、物理科学、科学共同体与社会、社会科学、人文科学、商务贸易、弃用的九大类主题。Springer Nature 为主题数据模型提供了 16 个属性，将 16 个属性分为类型、标签、标识符、参照、SKOS 表示 5 类。使用 rdf:type 表示类型；使用 rdfs:label 表示标签；标识符属性为主题的 Id 号；参照类属性包括 Springer Nature 主题词之间的相关关系、替代关系；SKOS 表示是通过 SKOS 命名空间对 Springer Nature 主题词进行组织表示，包括其首选词、非首选词、定义、注释、范围注释、主题词所属主题词表、上下位关系和族首词。

在各属性的取值中，与 Springer Nature 提供的相关主题词、替代主题词、被替代主题词对应的实例 URI 进行内部链接；并通过 SKOS 对 Springer Nature 的主题词进行组织，实现语义表示；其他属性取值均为文本值，如图 5 所示：

chinaXiv:201801.00105v1

图 5 主题数据模型

5 RDF 实现

Springer Nature SciGraph 关联开放数据平台目前对外只提供了 N-Triples 格式的三元组，N-Triples 是用于 RDF 的一种更简单的序列化，一种面向行的格式。每个三元组必须写成一个独立行，它由主语说明符、谓语说明符以及宾语说明符组成，后面还有一个句号。如果它们有 URI，那么它们用尖括号将绝对 URI 引用括起来[27]。截取 Springer Nature SciGraph 提供的有关主题词“genomics”的一些 N-Triples 格式的代码，如图 6 所示：

```
<http://www.springernature.com/scigraph/things/subjects/genomics>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    http://www.springernature.com/scigraph/ontologies/core/Subject.

<http://www.springernature.com/scigraph/things/subjects/genomics>
    <http://www.w3.org/2000/01/rdf-schema#label>    Genomics@en .

<http://www.springernature.com/scigraph/things/subjects/genomics>
    <http://www.springernature.com/scigraph/ontologies/core/id> genomics

<http://www.springernature.com/scigraph/things/subjects/genomics>
    <http://www.w3.org/2000/01/rdf-schema#isDefinedBy>
    <http://www.springernature.com/scigraph/things/subjects/>.

<http://www.springernature.com/scigraph/things/subjects/genomics>
    <http://www.springernature.com/scigraph/ontologies/core/webpage>
    http://www.nature.com/subjects/genomics.
```

图 6 “genomics”的一些 N-Triples 格式的关联数据

6 授权许可

授权许可通过明确发布和使用关联数据过程中的各项法律问题，包括所有权、发布权、使用权、收益权等，针对不同对象限定不同的权限，构建合理利用关联数据的保护机制，以达到促进数据开放、保证开放数据安全、提高数据重用性的目的。授开放许可是语义网环境下关联数据能够真正开放并长久发展的必要条件。授权许可通过嵌入数据，使得用户在访问数据时无需征得数据发布者的许可，授权许可包含了用户被许可的所有操作和未被许可的的操作。目前已有的授权模型有:免费文献许可协议(GNU free documentation license)、共享文件许可协议(common documentation license)、知识共享许可协议(creative commons license)等，各个协议能够实现对不同类型开放数据的保护。

Springer Nature SciGraph 发布关联数据的目的是为了将其科研数据融入关联数据网络中，使其在公共领域发挥作用，因而选择较为通用的知识共享许可协议作为授权许可协议 Springer Nature SciGraph 中的数据是在署名-非商业性使用国际 4.0 (CC BY-NC 4.0)授权模型下获取的[28]，表示允许在任何媒介以任何形式复制、发行本作品，并允许修改、转换或以本作品为基础进行创作，但不能将本作品用于商业目的[29]。通过 dcterms: license 属性表示该知识共享协议，属性值为 https://creativecommons.org/licenses/by-nc/4.0/表示对数据使用进

chinaXiv:201801.00105v1

行授权。

7 结语

Springer Nature SciGraph 关联开放数据平台的推出，是出版领域实现关联开放科研的起点。Springer Nature SciGraph 关联开放数据的实现，打破了原有数据组织结构体系，实现了数据关联、互操作、数据挖掘等功能，从概念角度对出版物进行描述，通过对 Springer Nature SciGraph 关联开放数据模型的分析，可知出版社作为发布出版物的源头，对出版进行语义描述对于实现数据互操作、数据关联有着重要的意义。我国出版领域应借鉴这种科研出版物实现关联开放模式，通过详细分析中文出版物包含的实体、属性及彼此之间的关系，选择合适的本体或在面对我国特有的文献（古籍、拓片等）时构建自己的本体，构建数据模型来实现中文出版物的语义描述和语义出版，同时指定授权许可明确数据使用过程中的各项法律问题。当然实现过程离不开软件平台的使用，考虑到成本问题和不同出版社之间数据的互操作问题，针对不同的学科，考虑使用开源软件和同一学科下的出版社之间采用统一的数据模型、统一的词汇表来实现出版物的关联开放。

作者贡献说明：

白林林： 负责论文的数据获取、提纲与撰写；

祝忠明： 负责论文的修订。

参考文献：

chinaXiv:201801.00105v1

ChinaXiv合作期刊

1[]LEHMANN J,AUER S,CAPADISLI S,et al.LDOW2017: 10th workshop on linked data on the Web [EB/OL]. [2017-05-25]. http://events.linkeddata.org/ldow2017/ldow-10th-workshop.pdf.
2[] From ScholarlyData.org to pioneering the linked open research cloud [EB/OL]. [2017-05-25]. http://aims.fao.org/ar/activity/blog/scholarlydataorg-pioneering-linked-open-research-cloud.
3[] Linked Research [EB/OL]. [2017-05-25]. https://linkedresearch.org/.
4[] Welcome to scholarlydata.org [EB/OL]. [2017-05-25]. http://www.scholarlydata.org/.
5[]The conference ontology [EB/OL]. [2017-05-25].http://www.scholarlydata.org/ontology/doc/.
6[]About RDA [EB/OL].[2017-05-25]. https://www.rd-alliance.org/about-rda.
7[] Springer Nature SciGraph: Supporting open science and the wider understanding of research [EB/OL]. [2017-05-25]. http://www.springernature.com/cn/group/media/press-releases/springer-nature-scigraph--supporting-open-science-and-the-wider-understanding-of-research/12129614.
8[]SciGraph Dataset Downloads [EB/OL]. [2017-05-26]. https://github.com/springernature/scigraph/wiki#getting-started.
9[]支持开放科研，施普林格·自然推出关联数据平台 [EB/OL]. [2017-05-26]. SciGraph http://www.toutiao.com/a6397639332211818754/.
10[] Nature.com Ontologies [EB/OL].[2017-05-27]. https://www.nature.com/ontologies/models/core/.
11[] GRID - Global Research Identifier Database [EB/OL]. [2017-05-27]. https://www.grid.ac/downloads.
12[] Australian and New Zealand standard research classification: fields of research [EB/OL]. [2017-05-28]. https://vocabs.ands.org.au/anzsrc-for.
13[]Digital object identifier system [EB/OL].[2017-05-29]. http://www.doi.org/index.html.
14[]Vocabulary of interlinked datasets (VoID) [EB/OL].[2017-05-30]. http://vocab.deri.ie/void#.
15[]Bibliographic ontology [EB/OL].[2017-05-26]. http://purl.org/ontology/bibo.
16[]Definition of the CIDOC conceptual reference model [EB/OL].[2017-05-30]. http://www.cidoc-crm.org/html/5.0.4/cidoc-crm.html.
17[]DBpedia [EB/OL]. [2017-05-31]. http://wiki.dbpedia.org/.
18[]FaBiO, the FRBR-aligned bibliographic ontology [EB/OL].[2017-05-31]. http://www.sparontologies.net/ontologies/fabio/source.html.
19[]MeSHlinked data (beta) [EB/OL]. [2017-05-31]. https://id.nlm.nih.gov/mesh/.
20[] Open biomedical ontologies [EB/OL]. [2017-05-31]. https://en.wikipedia.org/wiki/Open_Biomedical_Ontologies.
21[]PRISM metadata [EB/OL].[2017-06-01]. https://www.idealliance.org/prism-metadata/.
22[]vCardontology - for describing people and organizations [EB/OL]. [2017-06-02]. https://www.w3.org/TR/vcard-rdf/.
23[]VIVO-Integrated semantic framework [EB/OL].[2017-06-02]. http://bioportal.bioontology.org/ontologies/VIVO-ISF?p=classes&conceptid=root.
24[]Entity data [EB/OL].[2017-06-02]. https://www.wikidata.org/wiki/Special:EntityData/.
25[] N-Triples [EB/OL]. [2017-06-02]. https://en.wikipedia.org/wiki/N-Triples
26[]scigraph/articles.ttl at master [EB/OL]. [2017-06-02]. https://github.com/springernature/scigraph/blob/master/shapes/articles.ttl.
27[]RDF 1.1 N-Triples [EB/OL].[2017-06-04]. https://www.w3.org/TR/n-triples/.
28[] Creative commons — attribution-NonCommercial 4.0 International — CC BY-NC 4.0[EB/OL]. [2017-06-04]. https://creativecommons.org/licenses/by-nc/4.0/.
29[] Creative commons — 署名-非商业性使用 4.0 国际 — CC BY-NC 4.0 [EB/OL]. [2017-06-04]. https://creativecommons.org/licenses/by-nc/4.0/deed.zh.

Analysis of Springer Nature SciGraphLinked Open Data*

Bai Linlin[1,2]
[1]Lanzhou Library, Chinese Academy of Sciences, Lanzhou 730000
[2]University of Chinese Academy of Sciences, Beijing 100049

Abstract: ［**Purpose/significance**］The analysis of thelinked open data provided by the Springer Nature SciGraph platform provides a reference for the domestic publishers to use the linked data to promote the practice of linked open research in the scholarly communication and the semantic publishing, and push forward the further development of the open scientific research movement.**[Method/process**］This paper analyzed the entity objects, vocabularies and data models of the Springer Nature SciGraph platformin detail. ［**Result/conclusion**］Springer Nature SciGraph represents the data by building its own ontology and using a simpler serialized format N-Triples triple for RDF.As one of the world's largest publishers, Springer Nature Linked Data will provide some references for other publishers to realize the linked research in the future.

Keywords: Springer Nature SciGraphlinked open data open research

chinaXiv:201801.00105v1